# 2014 11th IAPR International Workshop on Document Analysis Systems

# (DAS 2014)

## Tours, France
## 7 – 10 April 2014

IEEE

**IEEE Catalog Number:** CFP1470E-POD
**ISBN:** 978-1-4799-3245-0

# 2014 11th IAPR International Workshop on Document Analysis Systems

# DAS 2014

## Table of Contents

**Message from the General Chairs and Program Chairs** ..... xi

**Committees** ..... xiii

**Program Committee and Subreviewers** ..... xvi

**Sponsors** ..... xx

---

### Oral-1: Document Classification and Retrieval

Plane Geometry Figure Retrieval with Bag of Shapes ..... 1
*Lu Liu, Xiaoqing Lu, Keqiang Li, Jingwei Qu, Liangcai Gao, and Zhi Tang*

Multiscale Stroke-Based Page Segmentation Approach ..... 6
*Mehdi Felhi, Salvatore Tabbone, and Maria V. Ortiz Segovia*

Business Forms Classification Using Earth Mover's Distance ..... 11
*Syed Saqib Bukhari, Markus Ebbecke, and Michael Gillmann*

### Oral-2: Writer Identification and Verification

Multiclass Learning for Writer Identification Using Error-Correcting Codes ..... 16
*Utkarsh Porwal, Chetan Ramaiah, Ashish Kumar, and Venu Govindaraju*

A Hierarchical Framework for Accent Based Writer Identification ..... 21
*Chetan Ramaiah and Venu Govindaraju*

Feasibility Study of Visualizing Diversity of Japanese Hiragana Handwritings by Multidimensional Scaling of Earth Mover's Distance toward Assisting Forensic Experts in Writer Verification ..... 26
*Yoshinori Akao, Atsushi Yamamoto, and Yoshiyasu Higashikawa*

## Oral-3: Text Detection and Correction

Flexible Noisy Text Correction ..... 31
*Andrey Sariev, Vladislav Nenchev, Stefan Gerdjikov, Petar Mitankin, Hristo Ganchev, Stoyan Mihov, and Tinko Tinchev*

Planting, Growing, and Pruning Trees: Connected Filters Applied to Document Image Analysis ..... 36
*Guillaume Lazzara, Thierry Géraud, and Roland Levillain*

Text Detection Using Delaunay Triangulation in Video Sequence ..... 41
*Liang Wu, Palaiahnakote Shivakumara, Tong Lu, and Chew Lim Tan*

A Typed and Handwritten Text Block Segmentation System for Heterogeneous and Complex Documents ..... 46
*P. Barlas, S. Adam, C. Chatelain, and T. Paquet*

## Oral-4: Handwritten Text Recognition and Graph Edit Distances

A System for Recognizing Online Handwritten Mathematical Expressions and Improvement of Structure Analysis ..... 51
*Anh Duc Le, Truyen Van Phan, and Masaki Nakagawa*

Curriculum Learning for Handwritten Text Line Recognition ..... 56
*Jérôme Louradour and Christopher Kermorvant*

Iterative Bipartite Graph Edit Distance Approximation ..... 61
*Kaspar Riesen, Rolf Dornberger, and Horst Bunke*

## Poster-1

A Context Based Text Summarization System ..... 66
*Rafael Ferreira, Frederico Freitas, Luciano de Souza Cabral, Rafael Dueire Lins, Rinaldo Lima, Gabriel França, Steven J. Simske, and Luciano Favaro*

A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents ..... 71
*Andreas Fischer, Micheal Baechler, Angelika Garz, Marcus Liwicki, and Rolf Ingold*

Local Binary Patterns for Arabic Optical Font Recognition ..... 76
*Anguelos Nicolaou, Fouad Slimane, Volker Maergner, and Marcus Liwicki*

NIST 2013 Open Handwriting Recognition and Translation (Open HaRT'13) Evaluation ..... 81
*Audrey Tong, Mark Przybocki, Volker Märgner, and Haikal El Abed*

Holistic Recognition of Online Handwritten Words Based on an Ensemble of SVM Classifiers ..... 86
*A. Srimany, S. Dutta Chowdhuri, U. Bhattacharya, and S.K. Parui*

Recognition of Handwritten Mathematical Characters on Whiteboards Using Colour Images ..... 91
*Behrang Sabeghi Saroui and Volker Sorge*

Historical Chinese Character Recognition Method Based on Style Transfer Mapping ..... 96
*Bohan Li, Liangrui Peng, and Jingning Ji*

Sequential Word Spotting in Historical Handwritten Documents ..... 101
*David Fernández-Mota, R. Manmatha, Alicia Fornés, and Josep Lladós*

Evaluation of Texture Features for Offline Arabic Writer Identification ..... 106
*Chawki Djeddi, Labiba-Souici Meslati, Imran Siddiqi, Abdelllatif Ennaji, Haikal El Abed, and Abdeljalil Gattal*

The RWTH Large Vocabulary Arabic Handwriting Recognition System ..... 111
*Mahdi Hamdani, Patrick Doetsch, Michał Kozielski, Amr El-Desoky Mousa, and Hermann Ney*

Context-Dependent Confusions Rules for Building Error Model Using Weighted Finite State Transducers for OCR Post-Processing ..... 116
*Mayce Al Azawi and Thomas M. Breuel*

Multilingual Off-Line Handwriting Recognition in Real-World Images ..... 121
*Michał Kozielski, Patrick Doetsch, Mahdi Hamdani, and Hermann Ney*

Multi-oriented Handwritten Annotations Extraction from Scanned Documents ..... 126
*Mohamed Benjlaiel, Rémy Mullot, and Adel M. Alimi*

The Influence of Language Orthographic Characteristics on Digital Word Recognition ..... 131
*Ofer Biller, Jihad El-Sana, and Klara Kedem*

Multi-oriented Text Recognition in Graphical Documents Using HMM ..... 136
*Partha Pratim Roy, Sangheeta Roy, and Umapada Pal*

Towards a Robust OCR System for Indic Scripts ..... 141
*Praveen Krishnan, Naveen Sankaran, Ajeet Kumar Singh, and C.V. Jawahar*

Printer Identification Using Supervised Learning for Document Forgery Detection ..... 146
*Sara Elkasrawi and Faisal Shafait*

Text Classification via iVector Based Feature Representation ..... 151
*Shengxin Zha, Xujun Peng, Huaigu Cao, Xiandan Zhuang, Pradeep Natarajan, and Prem Natarajan*

Spotting Symbol Using Sparsity over Learned Dictionary of Local Descriptors ..... 156
*Thanh-Ha Do, Salvatore Tabbone, and Oriol Ramos Terrades*

The A2iA Arabic Handwritten Text Recognition System at the Open HaRT2013 Evaluation ..... 161
*Théodore Bluche, Jérôme Louradour, Maxime Knibbe, Bastien Moysset, Mohamed Faouzi Benzeghiba, and Christopher Kermorvant*

A Cache Language Model for Whole Document Handwriting Recognition ..... 166
*Volkmar Frinken, Dimosthenis Karatzas, and Andreas Fischer*

On-line Handwritten Mathematical Expression Recognition Method Based on Statistical and Semantic Analysis ....171
*Yang Hu, Liangrui Peng, and Yejun Tang*

Word-Graph Based Handwriting Key-Word Spotting: Impact of Word-Graph Size on Performance ....176
*Alejandro Héctor Toselli and Enrique Vidal*

## Oral-5a: OCR, Word Spotting, and Recognition

Combining Focus Measure Operators to Predict OCR Accuracy in Mobile-Captured Document Images ....181
*Marçal Rusiñol, Joseph Chazalon, and Jean-Marc Ogier*

End-to-End Text Recognition Using Local Ternary Patterns, MSER and Deep Convolutional Nets ....186
*Michael Opitz, Markus Diem, Stefan Fiel, Florian Kleber, and Robert Sablatnig*

Adapting Tesseract for Complex Scripts: An Example for Urdu Nastalique ....191
*Qurat ul Ain Akram, Sarmad Hussain, Aneta Niazi, Umair Anjum, and Faheem Irfan*

Performance Improvement in Local Feature Based Camera-Captured Character Recognition ....196
*Takahiro Matsuda, Masakazu Iwamura, and Koichi Kise*

## Oral-5b: OCR, Word Spotting, and Recognition

OCR Performance Prediction Using a Bag of Allographs and Support Vector Regression ....202
*Tapan Kumar Bhowmik, Thierry Paquet, and Nicolas Ragot*

A Novel Learning-Free Word Spotting Approach Based on Graph Representation ....207
*Peng Wang, Véronique Eglin, Christophe Garcia, Christine Largeron, Josep Lladós, and Alicia Fornés*

Over-Generative Finite State Transducer N-Gram for Out-of-Vocabulary Word Recognition ....212
*Ronaldo Messina and Christopher Kermorvant*

## Oral-6: Super-Resolution and Born Digital Document Processing

CERMINE — Automatic Extraction of Metadata and References from Scientific Literature ....217
*Dominika Tkaczyk, Pawel Szostek, Piotr Jan Dendek, Mateusz Fedoryszak, and Lukasz Bolikowski*

End-to-End Conversion of HTML Tables for Populating a Relational Database ....222
*George Nagy, Sharad Seth, and David W. Embley*

Efficient Example-Based Super-Resolution of Single Text Images Based on Selective Patch Processing ..... 227
*Nibal Nayef, Joseph Chazalon, Petra Gomez-Krämer, and Jean-Marc Ogier*

Real-Time Document Image Super-Resolution by Fast Matting ..... 232
*Yun Zheng, Xudong Kang, Shutao Li, Yuan He, and Jun Sun*

## Oral-7: Ground Truthing and Performance Evaluation

Ground-Truth Production in the Transcriptorium Project ..... 237
*Basilis Gatos, Georgios Louloudis, Tim Causer, Kris Grint, Verónica Romero, Joan Andreu Sánchez, Alejandro H. Toselli, and Enrique Vidal*

An On-line Platform for Ground Truthing and Performance Evaluation of Text Extraction Systems ..... 242
*Dimosthenis Karatzas, Sergi Robles, and Lluis Gomez*

Ground-Truth and Performance Evaluation for Page Layout Analysis of Born-Digital Documents ..... 247
*Xin Tao, Zhi Tang, Canhui Xu, and Liangcai Gao*

## Poster-2

Forgery Detection Based on Intrinsic Document Contents ..... 252
*Amr Gamal Hamed Ahmed and Faisal Shafait*

Newspaper Article Extraction Using Hierarchical Fixed Point Model ..... 257
*Anukriti Bansal, Santanu Chaudhury, Sumantra Dutta Roy, and J.B. Srivastava*

A Seed-Based Segmentation Method for Scene Text Extraction ..... 262
*Bo Bai, Fei Yin, and Cheng Lin Liu*

Color Descriptor for Content-Based Drawing Retrieval ..... 267
*Christophe Rigaud, Dimosthenis Karatzas, Jean-Christophe Burie, and Jean-Marc Ogier*

A New One-Class Classification Method Based on Symbolic Representation: Application to Document Classification ..... 272
*Fahimeh Alaei, Nathalie Girard, Sabine Barrat, and Jean-Yves Ramel*

Automatic Training Set Generation for Better Historic Document Transcription and Compression ..... 277
*Gabriel de França Pereira e Silva, Rafael Dueire Lins, and Cesar Gomes*

Graph Model Optimization Based Historical Chinese Character Segmentation Method ..... 282
*Jingning Ji, Liangrui Peng, and Bohan Li*

Empirical Evaluation of CRF-Based Bibliography Extraction from Reference Strings ..... 287
*Manabu Ohta, Daiki Arauchi, Atsuhiro Takasu, and Jun Adachi*

Robustness Assessment of Texture Features for the Segmentation of Ancient Documents ..... 293
*Maroua Mehri, Van Cuong Kieu, Mohamed Mhiri, Pierre Héroux, Petra Gomez-Krämer, Mohamed Ali Mahjoub, and Rémy Mullot*

Fast and Optimal Binary Template Matching Application to Manga Copyright Protection ..... 298
*Mathieu Delalandre, Motoi Iwata, and Koichi Kise*

AreCAPTCHA: Outsourcing Arabic Text Digitization to Native Speakers ..... 304
*Menna Bakry, Mohamed Khamis, and Slim Abdennadher*

A Study to Achieve Manga Character Retrieval Method for Manga Images ..... 309
*Motoi Iwata, Atsushi Ito, and Koichi Kise*

Improving Classification of an Industrial Document Image Database by Combining Visual and Textual Features ..... 314
*Olivier Augereau, Nicholas Journet, Anne Vialard, and Jean-Philippe Domenger*

The Robustness of a New 3D CAPTCHA ..... 319
*Qi Ye, Youbin Chen, and Bin Zhu*

A Complete Logo Detection/Recognition System for Document Images ..... 324
*Alireza Alaei and Mathieu Delalandre*

A Two Level Algorithm for Text Detection in Natural Scene Images ..... 329
*Li Rong, Wang Suyu, and Zhixin Shi*

Land Map Images Binarization Based on Distance Transform and Adaptive Threshold ..... 334
*Samit Biswas, Sekhar Mandal, Amit Kumar Das, and Bhabatosh Chanda*

A New Laplacian Method for Arbitrarily-Oriented Word Segmentation in Video ..... 339
*P. Shivakumara, M. Suhil, D.S. Guru, and C.L. Tan*

Separation of Graphics (Superimposed) and Scene Text in Video Frames ..... 344
*P. Shivakumara, N.V. Kumar, D.S. Guru, and C.L. Tan*

The Maurdor Project: Improving Automatic Processing of Digital Documents ..... 349
*Sylvie Brunessaux, Patrick Giroux, Bruno Grilhères, Mathieu Manta, Maylis Bodin, Khalid Choukri, Olivier Galibert, and Juliette Kahn*

Design of Unsupervised Feature Extraction System for On-line Bangla Handwriting Recognition ..... 355
*Volkmar Frinken, Nilanjana Bhattacharya, and Umapada Pal*

Logical Labeling of Fixed Layout PDF Documents Using Multiple Contexts ..... 360
*Xin Tao, Zhi Tang, Canhui Xu, and Yongtao Wang*

Text Line Segmentation Based on Matched Filtering and Top-Down Grouping for Handwritten Documents ..... 365
*Youbao Tang, Xiangqian Wu, and Wei Bu*

**Author Index** ..... 370